THE USE OF CHI-SQUARE STATISTICS FOR CATEGORICAL DATA PROBLEMS

by

Stephen E. Fienberg

Technical Report #313

March 15, 1978

Department of Applied Statistics
School of Statistics
University of Minnesota

ABSTRACT

The use of chi-squared statistics for categorical data problems was initiated by Karl Pearson, but it took several years before the asymptotic distribution of these statistics was well understood. The general structure of asymptotic results for chi-squared statistics is reviewed and the applicability of the general structure to a variety of problems of practical interest is discussed. These problems include the use of chi-squared statistics in small-sample situations and in large sparse tables, in cluster sampling, and in cases where they do not have asymptotic chi-square distributions.

Key Words: Categorical data; Chi-square statistics; Cluster sampling; Goodness of fit; Large sparse multinomials; Likelihood ratio statistics; Multinomial sampling model; Small sample properties.

1. Introduction

During the 1950's several review papers on the use of chi-square goodness-of-fit statistics appeared, and most of these had advice on the use of these statistics with their related asymptotic chi-square distributions (e.g. see Cochran, 1952, 1954). Other papers dealt directly with the theory and gave the asymptotic distribution of chi-square statistics computed in a variety of different ways (e.g. see Watson, 1959). In the intervening years considerable attention has been focussed on the development of methods for the analysis of categorical data, primarily through the use of loglinear models (e.g. see Bishop, Fienberg, and Holland, 1975, or Plackett, 1974). The expanding interest in this topic has kindled further efforts on both the theory for chi-square tests and their use in statistical practice. The present paper provides a review of some of this recent work.

In Section 2 we summarize the general structure for results on the asymptotic distribution of chi-square statistics for categorical data problems generated by multinomial sampling schemes. The results for product-multinomial sampling schemes are quite similar and are omitted for this reason. The focus of these results is on the behavior of the test statistics under composite null hypotheses and there is no discussion of asymptotic distributions under alternative hypotheses, and of related power considerations. This general asymptotic structure is well-known and we include it here so that we can refer to specific results at crucial junctures later in the paper.

Section 3 then describes some results on the use of the standard asymptotic chi-square reference distributions in problems with small

sample sizes. We give particular attention to the comparisons between the usual Pearson and likelihood ratio statistics. This section concludes with a description of an "improved" likelihood ratio test.

In Sections 4 and 5 we describe some asymptotic results for goodness-of-fit statistics in nonstandard problems. Section 4 deals with problems involving complex estimation of parameters, and Section 5 considers the distribution of the usual test statistics when the data are generated by cluster sampling.

Finally in Section 6 we sketch the asymptotic framework for large sparse nultinomial structures, where the sample size and the number of cells both get large at the same rate. These results provide another way to think about small-sample properties of test-statistics in categorical data problems with a relatively large number of cells.

2. The Asymptotic Machinery for Chi-square Theorems

The core asymptotic results on the chi-square distribution of the Pearson and loglikelihood ratio goodness-of-fit statistics are by now well-known. In this section we establish the basic notation for this paper and briefly summarize the basic theorems in a form to allow their adaptation to non-standard situations. For a more detailed development the interested reader is referred to Bishop, Fienberg, and Holland (1975).

Let $\underset{\sim}{X} = (X_1, X_2, \ldots, X_t)$ have the multinomial distribution $\mathcal{M}(N, \underset{\sim}{\pi})$ where $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_t)$. Then

$$E(\underset{\sim}{X}) = N\underset{\sim}{\pi} \quad , \tag{2.1}$$

and

$$\text{Cov}(\underset{\sim}{X}) = N(\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi}) \tag{2.2}$$

where $\underset{\sim}{D}_{\underset{\sim}{\pi}}$ is the diagonal matrix based on $\underset{\sim}{\pi}$. Let $\hat{\underset{\sim}{p}} = N^{-1}\underset{\sim}{X}$ be the vector of sample proportions. Then

Theorem 1. As $N \to \infty$, the random vector $\sqrt{N}(\hat{\underset{\sim}{p}} - \underset{\sim}{\pi})$ converges in distribution to a multivariate normal with mean $\underset{\sim}{0}$ and covariance matrix $\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi}$.

Let $\mathcal{S}_t$ be the t dimensional probability simplex

$$\mathcal{S}_t = \{\underset{\sim}{p}: \; p_i \geqq 0 \text{ and } \sum_{i=1}^{t} p_i = 1\} \quad . \tag{2.3}$$

Both the vector of true cell probabilities $\underset{\sim}{\pi}$ and the vector of observed proportions $\hat{\underset{\sim}{p}}$ are points in $\mathcal{S}_t$. In the typical categorical data problem $\underset{\sim}{\pi}$ is unknown, and is assumed to be a function of a reduced number of parameters, denoted by the vector $\underset{\sim}{\varphi}$, which are also unknown. We typically

assume that $\underset{\sim}{\pi}$ lies in some subspace of $\mathcal{S}_t$, denoted by M (for model) and characterized by a vector of parameters $\underset{\sim}{\theta}$ of dimension s. Since $\underset{\sim}{\pi}$ is a function of $\underset{\sim}{\theta}$ we write $\underset{\sim}{\pi} = \underset{\sim}{f}(\underset{\sim}{\theta})$. If the model M is correct then $\underset{\sim}{f}(\underset{\sim}{\varphi})$ lies in M, otherwise it does not.

The first task in assessing the goodness-of-fit of the model M is to estimate the vector $\underset{\sim}{\theta}$. We can do this by the method of maximum likelihood estimation or any other asymptotically efficient method. Let $\hat{\underset{\sim}{\theta}}$ be the maximum likelihood estimate (MLE) of $\underset{\sim}{\theta}$.

Theorem 2. Assume that $\underset{\sim}{\pi} = \underset{\sim}{f}(\underset{\sim}{\varphi})$, i.e. that $\underset{\sim}{f}(\underset{\sim}{\varphi})$ lies in M. Then under suitable regularity conditions

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{\varphi} + (\hat{\underset{\sim}{p}} - \underset{\sim}{\pi})\underset{\sim}{D}_{\underset{\sim}{\pi}}^{-\frac{1}{2}}\underset{\sim}{A}(\underset{\sim}{A}'\underset{\sim}{A})^{-1} + 0_p(N^{-\frac{1}{2}}) \tag{2.4}$$

where $\underset{\sim}{A}$ is a $t \times s$ matrix of rank s whose (i,j) element is

$$a_{ij} = \pi_i^{-\frac{1}{2}}\left(\frac{\partial f_i(\underset{\sim}{\varphi})}{\partial \theta_j}\right) \tag{2.5}$$

An important consequence of Theorem 2 is the asymptotic distribution of $\hat{\underset{\sim}{\theta}}$ under the hypothesis that the model is correct.

Theorem 3. Under the conditions of Theorem 2, the asymptotic distribution of $\sqrt{N}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\varphi})$ is

$$\mathcal{N}(\underset{\sim}{0}, (\underset{\sim}{A}'\underset{\sim}{A})^{-1}) \quad . \tag{2.6}$$

Once we have the MLE $\hat{\underset{\sim}{\theta}}$ and thus $\hat{\underset{\sim}{\pi}} = \underset{\sim}{\pi}(\hat{\underset{\sim}{\theta}})$ we can assess the goodness-of-fit of M by means of one of the standard statistics, such as

loglikelihood ratio

$$G^2 = 2N\sum_i \hat{p}_i \log\left(\frac{\hat{p}_i}{\hat{\pi}_i}\right) \quad , \tag{2.7}$$

Pearson

$$X^2 = N \sum_i \frac{(\hat{p}_i - \hat{\pi}_i)^2}{\hat{\pi}_i} , \tag{2.8}$$

Freeman-Tukey

$$F^2 = 4N \sum_i (\sqrt{\hat{p}_i} - \sqrt{\hat{\pi}_i})^2 . \tag{2.9}$$

A crucial result regarding the asymptotic equivalence of these statistics, assuming that the model M is correct, is as follows.

Theorem 4. Let $\hat{\underset{\sim}{\pi}}$ be any estimate of $\underset{\sim}{\pi}$ $(\pi_i > 0)$ such that $\hat{\underset{\sim}{p}}$ and $\hat{\underset{\sim}{\pi}}$ have a joint limiting normal distribution, i.e. the limiting distribution of $\sqrt{N}((\hat{\underset{\sim}{p}},\hat{\underset{\sim}{\pi}}) - (\underset{\sim}{\pi},\underset{\sim}{\pi}))$ is $\mathcal{N}(\underset{\sim}{0},\underset{\sim}{\Sigma})$ for some covariance matrix $\underset{\sim}{\Sigma}$. Then G^2, X^2, and F^2 all have the same limiting distribution.

To complete the general asymptotic machinery we need to determine the distribution of X^2 under M for the estimator $\hat{\underset{\sim}{\pi}}$ of $\underset{\sim}{\pi}$ from Theorem 4. Using standard results on quadratic forms of multivariate normal random vectors, we have:

Theorem 5. Under the conditions of Theorem 4, as $N \to \infty$ the distribution of X^2 converges to $\sum_{i=1}^{t-1} \lambda_i z_i^2$, where z_i^2 are independent chi-square variables with one degree of freedom and the λ_i are the nonzero eigenvalues of

$$\underset{\sim}{D}_{\underset{\sim}{\pi}}^{-\frac{1}{2}} (\underset{\sim}{\Sigma}_{11} - \underset{\sim}{\Sigma}_{12} - \underset{\sim}{\Sigma}_{21} + \underset{\sim}{\Sigma}_{22}) \underset{\sim}{D}_{\underset{\sim}{\pi}}^{-\frac{1}{2}} , \tag{2.10}$$

with $\underset{\sim}{\Sigma}$ partitioning according to $(\hat{\underset{\sim}{p}},\hat{\underset{\sim}{\pi}})$:

$$\underset{\sim}{\Sigma} = \begin{pmatrix} \underset{\sim}{\Sigma}_{11} & \underset{\sim}{\Sigma}_{12} \\ \underset{\sim}{\Sigma}_{21} & \underset{\sim}{\Sigma}_{22} \end{pmatrix} .$$

When $\hat{\underset{\sim}{\pi}}$ is the MLE of $\underset{\sim}{\pi}$, based on the multinomial vector $\underset{\sim}{x}$, the matrix $\underset{\sim}{\Sigma}$ has the form

$$\underset{\sim}{\Sigma} = \begin{pmatrix} \underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi} & (\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi})\underset{\sim}{L} \\ \underset{\sim}{L}'(\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi}) & \underset{\sim}{L}'(\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi})\underset{\sim}{L} \end{pmatrix} \tag{2.11}$$

where

$$\underset{\sim}{L} = \underset{\sim}{D}_{\underset{\sim}{\pi}}^{-\frac{1}{2}}\underset{\sim}{A}(\underset{\sim}{A}'\underset{\sim}{A})^{-1}\underset{\sim}{A}'\underset{\sim}{D}_{\underset{\sim}{\pi}}^{\frac{1}{2}} \quad . \tag{2.12}$$

The asymptotic chi-square distribution of X^2 then follows from Theorem 5.

Theorem 6. Under the conditions of Theorem 2, if $\hat{\underset{\sim}{\pi}} = \underset{\sim}{\pi}(\hat{\underset{\sim}{\theta}})$ where $\underset{\sim}{\theta}$ is estimated by maximum likelihood, then the limiting distribution of X^2 as $N \to \infty$ is χ^2_{t-s-1}.

Since X^2, G^2, and F^2 all have the same limiting distribution when $\underset{\sim}{f}(\underset{\sim}{\varphi})$ lies in M, then each has an asymptotic chi-square distribution with t-s-1 degrees of freedom when the model M is correct.

3. Small-sample Properties of Chi-square Statistics

3.1. Monte Carlo Studies

It is all well and good to know the asymptotic distribution of G^2, X^2, and F^2 under the hypothesis that $\underline{\pi}(\underline{\omega})$ lies in M, but these asymptotic results tell us little about when the chi-square approximations can be viewed as adequate. In the pre-computer era advice by such distinguished statisticians as Cochran and Fisher was based on practical experience and intuition, and led to standard adequacy rules such as: "the minimal expected cell size should exceed 5". Such rules tended to be somewhat conservative, and more recent Monte Carlo studies by Odoroff (1970), Yarnold (1970), and Larntz (1978) suggest that, at least for tests conducted at a nominal 0.05 level of significance, the goodness-of-fit statistics often achieve the desired level when minimum expected cell values are approximately 1.0. In this section we briefly review some of the results of the most recent of these Monte Carlo studies by Larntz (1978), and his recommendations.

Larntz (1978) looked at a variety of categorical data problems with estimated parameters, and he determined the exact levels for three goodness-of-fit statistics and varying sample sizes using a nominal 0.05 level test. He explored the small sample properties of G^2, X^2, and

$$T^2 = \sum_i \left(\sqrt{N\hat{p}_i} + \sqrt{N\hat{p}_i + 1} - \sqrt{4N\hat{\pi}_i + 1}\right)^2 \qquad (3.1)$$

which is a variant of the Freeman-Tukey statistic (2.9) with somewhat more stable small sample properties.

Illustrative of his Monte Carlo results are those in Table 1 for a test of no second-order interaction in a $3 \times 3 \times 3$ contingency table ($u_{123(ijk)} = 0$ for all i, j, k in the notation of Bishop, Fienberg, and

Holland, 1975). Larntz generated 2000 random trials for each of 4 versions of the model of complete independence ($u_{12} = u_{13} = u_{23} = u_{123} = 0$), and then tested for no second-order interaction using G^2, X^2, and T^2. To handle sample zeros in the two-way marginals (the minimal sufficient statistics) he extended the MLE's in such cases by continuity, to provide well-defined procedures. In all but the sparsest of situations (e.g. N = 20 and N = 40) the small sample behavior of X^2 appears to be remarkably stable and the actual level comes quite close to the nominal 0.05 level. T^2 tends to be quite conservative for very small N ($p \ll 0.05$) and somewhat liberal for moderate N. The likelihood ratio statistic G^2 tends to reject substantially more often than is expected for moderate sample sizes (see the comparable results for two-way tables in Margolin and Light, 1974). What is especially remarkable here is that there are 27 cells, and so even for N = 100, the <u>average number of observations per cell is less than 3.5</u>!

Because of the somewhat abberant behavior of G^2 (and to a lesser extent T^2) for very small samples that was apparent in almost all of his Monte Carlo work, Larntz postulated that the discrepancy in behavior was due to the differing influences given to very small <u>observed</u> counts. He then calculated the minimum contribution to each of X^2, G^2, and T^2 resulting first from a single observed count of 0, and then from a single observed count of 1. For X^2 the contribution of a zero observed value is equal to the corresponding expected value, while for T^2 the contribution equals $(1 - \sqrt{4E + 1})^2$ where E is the expected value. For G^2, the contribution of a zero count appears to be zero but this is not actually the case since the zero then effects all other cells. In fact the minimum contribution of a zero count to G^2 occurs when the remaining counts are spread out evenly over the other cells in exact proportion to the expected values of those cells. Thus

minimum contribution

$$= \lim_{N\to\infty} 2N \log\left(\frac{N}{N-E}\right) = 2E \quad ,$$

and a zero count will have twice the impact on G^2 as on X^2. T^2 is not affected quite as much as G^2 for small values of the cell expectation, but as the latter grows in excess of 2.0 the effect of T^2 exceed the effect on G^2.

For multiple zero observed cells the minimum contribution to X^2 equals the sum of the corresponding expected values, and the minimum contribution to G^2 is twice that quantity. Similar results are true when there are one or more observed counts of 1, except the impact on T^2 tends to be less severe than in the case of observed zeros.

Larntz summarizes his results as follows:

1) Using as a criterion the closeness of the small sample distribution to the asymptotic chi-square approximation under the null hypothesis, X^2 is preferable to G^2 and T^2.

2) The relatively high type I error rates from G^2 and T^2 result from the large contributions to the chi-square value for very small observed counts in cells with moderate expected values.

3) Even when the minimum expected cell value in a table is between 1 and 4 in size, a P-value based on the asymptotic chi-square approximation is "on average" about right for X^2 but is understated for G^2 and T^2.

If one is concerned only about relative orders of magnitude of the P-values associated with goodness-of-fit tests such as X^2, G^2, and T^2, then Larntz's results suggest that values of N equal to four or five times the number of cells are adequate for the use of the asymptotic chi-square results. Keeping the ratio N/t roughly constant (say at 4 or 5)

leads, however, to a different type of asymptotics, which we discuss in Section 6. Obvious exceptions to the rule of "average cell sizes of 4 or 5" occur when most of the sample size is concentrated in a few cells with relatively large cell counts.

3.2. Improved Likelihood Ratio Tests

Williams (1976) has derived a modification to the likelihood ratio test wherein the statistic G^2 is multiplied by a scale factor chosen to make the moments of G^2 (under the null hypothesis) match those of the reference χ^2 distribution, ignoring terms of order N^{-2}. Williams' results pertain to loglinear models for complete multidimensional tables, with closed form MLE's, and uses independent Poissen variables as the sampling model for the cell frequencies.

This approach leads to replacing G^2 by

$$G^2_{adj} = \frac{1}{q} G^2 , \tag{3.2}$$

where the inverse of the multiplier, q, is given by

$$\begin{aligned} q = 1 + \frac{1}{6\nu} \text{ (sum of reciprocals of expected cell frequencies} \\ - \text{ sums of expectations of marginal frequencies in the numerators of the maximum likelihood estimators} \\ + \text{ sums of expectations of marginal frequencies in the denominators of the maximum likelihood estimators),} \end{aligned} \tag{3.3}$$

and ν is the degrees of freedom. In the case of independence in a two-dimensional $I \times J$ table, q takes the form:

$$q = 1 + \frac{1}{6(I-1)(J-1)} \left[\sum\sum_{ij} \frac{1}{m_{ij}} - \sum_i \frac{1}{m_{i+}} - \sum_j \frac{1}{m_{+j}} + \frac{1}{N}\right] \tag{3.4}$$

which takes as its minimum possible value

$$q_{min} = 1 + \frac{(I+1)(J+1)}{6N} \quad . \tag{3.5}$$

Williams suggests the use of q_{min} in (3.2), in part to avoid the problem of estimating m_{ij}^{-1} in (3.4).

Using the adjusted statistic in (3.2) in place of G^2 has the effect of reducing the size of the test statistic (since $q_{min} > 1$), and thus seems to be in accord with the small sample results of Section 3.1. There we noted that the actual small-sample p-values based on the chi-square approximation for G^2 were somewhat understated, and that this effect was somewhat more pronounced in the presence of sample zeros. To my knowledge, no one has directly examined the small sample properties of G^2_{adj} relative to the usual goodness-of-fit statistics.

Williams (1976) speculates that these results using a version of q_{min} extend to situations where the MLE's do not have closed form expressions.

4. The Use of Standard Asymptotics for Non-Standard Situations

Not all problems lead to the happy simplicity of the standard asymptotic chi-square result of Theorem 6. More often than not, in non-standard situations where the parameter vector $\underset{\sim}{\theta}$ (and thus $\underset{\sim}{\pi}$) is estimated in a complex way or from an alternative data source, we need to rely directly on Theorems 4 and 5. Thus the basic goodness-of-fit statistics have asymptotic distributions which are linear combinations of χ^2's, where some of the weights are different from zero or one.

Two examples of these non-standard results involve two independent multinomial samples, with the same underlying vector of cell probabilities, $\underset{\sim}{\pi}(\underset{\sim}{\theta})$, i.e.

$$\underset{\sim}{X} \sim \mathcal{M}(N, \underset{\sim}{\pi}(\underset{\sim}{\theta}))$$
$$\underset{\sim}{Y} \sim \mathcal{M}(N^*, \underset{\sim}{\pi}(\underset{\sim}{\theta})) \quad .$$

Now suppose both N and N^* tend to ∞ in such a way that

$$\lim \frac{N}{N^*} = \tau \quad . \tag{4.1}$$

If $\underset{\sim}{\theta}$ is estimated by $\underset{\sim}{\theta}^*$ (the MLE from the $\underset{\sim}{Y}$-sample), and X^2 is computed using the $\underset{\sim}{X}$-sample, then it follows from Theorems 4 and 5 that (see Chase, 1972)

$$X^2 \longrightarrow \sum_{i=1}^{t-s-1} z_i^2 + (1+\tau) \sum_{i=t-s}^{t-1} z_i^2 \quad , \tag{4.2}$$

where the z_i^2 are independent chi-squares with one degree of freedom, i.e. the asymptotic distribution is

$$\chi^2_{t-s-1} + (1+\tau)\chi^2_s \quad . \tag{4.3}$$

Note that this limiting distribution is stochastically larger than χ^2_{t-1}, which is the reference distribution for the situation where we test a predetermined hypothesis, $\underset{\sim}{\pi} = \underset{\sim}{\pi}_0$.

While this result may at first seem surprising it has a rather neat interpretation in terms of cross-validation. We can think of the $\underset{\sim}{Y}$-sample as being used for model selection and estimation, and the $\underset{\sim}{X}$-sample for (cross-) validation in terms of its fit to the model as estimated from the $\underset{\sim}{Y}$-sample. An optimal division of data between the $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ samples then corresponds to a choice of τ as close as possible to zero, for in that case the reference distribution of X^2 is approximately χ^2_{t-1}. Choosing τ close to zero means that we use almost all the data for fitting and model choice, i.e. N^* is large relative to N, and just a bit for validation. Because we do need more than 1 or 2 observations for validation, we cannot quite go to the "leave 1 (or 2) out" rules suggested by Stone (1974).

As an alternative to the scheme leading to (4.2), we might estimate $\underset{\sim}{\theta}$ by $\hat{\underset{\sim}{\theta}}^{**}$, the MLE from the pooled sample $\underset{\sim}{X} + \underset{\sim}{Y}$, but then compute X^2 using the observed counts from the $\underset{\sim}{X}$-sample alone. In this case Murthy and Gafarian (1970) have shown that

$$X^2 \longrightarrow \chi^2_{t-s-1} + \left(1 - \frac{\tau}{1+\tau}\right)\chi^2_s \quad . \tag{4.4}$$

Here the limiting distribution is bounded between the χ^2_{t-s-1} and χ^2_{t-1} distributions, and approaches the χ^2_{t-s-1} distribution as $\tau \to \infty$, rather than as $\tau \to 0$.

Extensions to these results are fairly direct, and rely on the basic asymptotic theory of Section 2. Larntz (1971) proved an extension of the Chase result, (4.2), to the situation of two independent samples with different cell probabilities vectors depending on

the same underlying parameters, $\underset{\sim}{\theta}$:

$$\underset{\sim}{X} \sim \mathcal{M}(N, \underset{\sim}{\pi}(\underset{\sim}{\theta}))$$

$$Y \sim \mathcal{M}(N^*, \underset{\sim}{\pi}^*(\underset{\sim}{\theta})).$$

Then the asymptotic distribution of X^2 using the observed values from the $\underset{\sim}{X}$-sample and $\hat{\underset{\sim}{\theta}}^*$ from the $\underset{\sim}{Y}$-sample is

$$\chi^2_{t-s-1} + \sum_{i=1}^{s} (1 + \tau\gamma_i)\, z_i^2. \qquad (4.5)$$

The γ_i's are the s nonzero eigenvalues of the matrix

$$\underset{\sim}{I} - \sqrt{\underset{\sim}{\pi}}'\sqrt{\underset{\sim}{\pi}} + \tau\underset{\sim}{A}_1' \,(\underset{\sim}{A}_2\underset{\sim}{A}_2')^{-1}\, \underset{\sim}{A}_1 \quad ,$$

where $\underset{\sim}{A}_1$ is defined by (2.5) using $\underset{\sim}{\pi} = \underset{\sim}{f}(\underset{\sim}{\varphi})$ and $\underset{\sim}{A}_2$ is similarly defined but using $\underset{\sim}{\pi}^* = \underset{\sim}{f}^*(\underset{\sim}{\varphi})$.

For Larntz's result the probability vectors for the two samples differ but the underlying parameters remained the same. Alternatively the probability vectors might have the same form but some of the components of $\underset{\sim}{\theta}$ may differ from sample to sample. For example, in a series of multi-dimensional contingency tables it might be reasonable to assume that there is no third or higher order interactions, that the second order interaction is the same for all tables, but that the two-dimensional marginal totals change from table to table. Then we might wish to estimate the MLE $\hat{\underset{\sim}{\pi}}_{\underset{\sim}{Y}}$ for $\underset{\sim}{\pi}_{\underset{\sim}{Y}}$ for one table and then adjust $\hat{\underset{\sim}{\pi}}_{\underset{\sim}{Y}}$ to have the marginal totals of a second table using iterative proportional fitting (see Bishop, Fienberg, and Holland 1975), yielding $\hat{\underset{\sim}{\pi}}^*_{\underset{\sim}{X}}$. Finally we would compare $\hat{\underset{\sim}{p}}_{\underset{\sim}{X}}$ with $\hat{\underset{\sim}{\pi}}^*_{\underset{\sim}{X}}$ using X^2. The situation here involves two independent multinomials:

$$\underset{\sim}{X} \sim \mathcal{M}(N, \underset{\sim}{\pi}(\underset{\sim}{\theta}^{(1)}, \underset{\sim}{\theta}^{(2)}))$$

$$\underset{\sim}{Y} \sim \mathcal{M}(N^*, \underset{\sim}{\pi}(\underset{\sim}{\theta}^{(1)}, \underset{\sim}{\theta}^{(2)*}))$$

where $\underset{\sim}{\theta}^{(2)}$ and $\underset{\sim}{\theta}^{(2)*}$ are not necessarily the same. To compute $\hat{\underset{\sim}{\pi}}_{\underset{\sim}{X}}$ we use an estimate of $\underset{\sim}{\theta}^{(1)}$ from the $\underset{\sim}{Y}$-sample, and an estimate of $\underset{\sim}{\theta}^{(2)}$ from the $\underset{\sim}{X}$-sample. (We could alternatively choose to estimate $\underset{\sim}{\theta}^{(1)}$ from the two samples simultaneously.) The asymptotic result in this case takes an almost predictable form based on Theorem 5, and Brier (1978) is currently exploring simplifications of the key matrix given by expression (2.10) for some special cases such as the example described above.

5. Asymptotic Distribution of X^2 for Cluster Sampling

A question often asked in practice is: How inappropriate is the use of the usual goodness-of-fit statistics and the standard chi-square reference distributions in situations where the data come from a cluster sample rather than the simple random sampling model implied by the multinomial distribution? In this section we answer this question by describing some recent results of Brier (19 8).

Brier considers the following situation. For each of r clusters, consider independent random samples of size n :

$$\underset{\sim}{X}^{(i)} \sim \mathcal{M}(n, \underset{\sim}{p}^{(i)}) \quad . \tag{5.1}$$

Suppose that the $\underset{\sim}{p}^{(i)}$ are themselves viewed as independent identically distributed random variables from a Dirichlet distribution with density

$$f(\underset{\sim}{p}|k,\underset{\sim}{\pi}) = \frac{\Gamma(K)}{\prod_{i=1}^{t} \Gamma(K\pi_i)} \prod_{i=1}^{t} p^{K\pi_i - 1} \quad , \tag{5.2}$$

where $\underset{\sim}{\pi}$ is a probability vector lying in $\mathcal{S}_t$ and $K > 0$. Then the marginal distribution of $\underset{\sim}{X}^{(i)}$ (integrating over $\underset{\sim}{p}^{(i)}$) is Dirichlet-multinomial (see, for example, Mosimann, 1962), and the r clusters can then be viewed as a random sample of size r from the DM $(n; \underset{\sim}{\pi}, K)$ distribution.

Plackett and Paul (1978a) discuss the Dirichlet-multinomial distribution in a somewhat different but closely related context, and they note that the cluster sampling models of Cohen (1976) and Altham (1976) can be viewed as special cases.

If we were to act as if we did not have the clusters, we would add across them and work with the data vector

$$\underset{\sim}{X} = \sum_{i=1}^{r} \underset{\sim}{X}^{(i)} \tag{5.3}$$

instead of the r data vectors for each cluster.

A more precise statement of the question posed in the opening paragraph of this section is: What is the asymptotic distribution of the Pearson X^2 statistic using the vector $\underset{\sim}{X}$ from (5.3) to test a hypothesis about $\underset{\sim}{\pi}$, as if there were no clustering? The answer again follows from the application of Theorems 4 and 5 of Section 2.

Note that we are not hampered by the replacement of multinomial by Dirichlet-multinomial sampling, since Theorem 4 only requires a joint limiting normal distribution for $N^{-1}\underset{\sim}{X}$ (where $N = rn$ is the overall sample size), and any estimate of $\underset{\sim}{\pi}$, $\hat{\underset{\sim}{\pi}}$. Then Theorem 5 makes use only of the mean and variance of the Dirichlet-Multinomial which have a remarkable resemblance to those of the multinomial with

$$E\,(\underset{\sim}{X}^{(i)}) = n\,\underset{\sim}{\pi} \tag{5.4}$$

and

$$\mathrm{Cov}\,(\underset{\sim}{X}^{(i)}) = Cn\,(\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}'\underset{\sim}{\pi}), \tag{5.5}$$

where

$$C = \frac{n + K}{1 + K}. \tag{5.6}$$

The net result of some detailed asymptotic manipulations is:

Theorem 7 (Brier, 1978). Suppose $\underset{\sim}{X}$ is defined by (5.3), and X^2 is computed as in (2.8). Then under the conditions of Theorem 2, if $\underset{\sim}{\pi} = \underset{\sim}{f}\,(\underset{\sim}{\theta})$ is estimated by maximum likelihood erroneously assuming a multinomial sampling model,

$$X^2 \longrightarrow C\, \chi^2_{t-s-1} \tag{5.7}$$

as $N \to \infty$.

Theorem 7 is closed related to results on inference sensitivity for Poisson mixtures derived by Plackett and Paul (1978b).

The scalar multiple C for the usual chi-square reference distribution in Theorem 7 is unfortunately unknown. Since $K > 0$ it is clear that

$$n > C \geq 1. \tag{5.8}$$

The upper limit corresponds to the bounds noted by Altham (1976) for the case $r = 3$. If the results for sample members of each cluster are highly interrelated then C will be close to n; if they are almost independent C will be near 1.

In order to make some practical use of this result Brier suggests computing the moment estimate

$$\hat{C} = \frac{1}{n(r-1)(t-1)} \sum_{i=1}^{r} (\underset{\sim}{X}^{(i)} - n\,\underset{\sim}{\tilde{\pi}})\, \underset{\sim}{D}^{-1}_{\underset{\sim}{\tilde{\pi}}}\, (\underset{\sim}{X}^{(i)} - n\,\underset{\sim}{\tilde{\pi}})' \tag{5.9}$$

where

$$\underset{\sim}{\tilde{\pi}} = N^{-1}\,\underset{\sim}{X} = (rn)^{-1} \sum_{i=1}^{r} \underset{\sim}{X}^{(i)}. \tag{5.10}$$

The estimate $\hat{C}$ is X^2-like in form and measures the homogeneity of the counts in the t categories across clusters. This estimate for C is consistent, and thus

$$\frac{1}{\hat{C}}\, X^2 \longrightarrow \chi^2_{t-s-1}. \tag{5.11}$$

If $\hat{C} < 1$, it seems reasonable to use X^2 without the scaling constant. Finney (1971) in a totally different context proposes the use of a

hetergeneity factor, in the form of a chi-square statistic divided by its degrees of freedom, to adjust the usual chi-square statistic. This can be viewed as a special case of the present result.

An alternative to using an "adjusted" version of the "usual" X^2 or G^2 statistics is to develop a direct approach for the $\binom{t+n-1}{n}$ possible outcomes for each of the r clusters and then applying the standard asymptotics for the resulting categorical structure. This approach is useful primarily for small t and n, and Brier (1978) has some small sample results suggesting that it does not work well even for moderate values of t. He also compares the two approaches in terms of power under near alternatives.

6. The Asymptotics of Large Sparse Multinomials

Much has been written about the desirability of collapsing in the presence of cells with counts of 0 and 1, and of the supposed lack of "information" in large sparse multidimensional tables. But the fact remains that with the extensive questionnaires of modern-day sample surveys, and the detailed and painstaking inventory of variables measured by biological and social scientists, the statistician is often faced with large sparse arrays, chock full of 0's and 1's, in need of careful analysis. Some recent analytical results suggest, however, that the beleaguered statistician need not despair: a new form of asymptotic structure provides the underpinnings for the analysis of large sparse multinomials, and the results for such asymptotics dovetail nicely with the "standard" small-sample results mentioned in Section 3.

The traditional asymptotic machinery outlined in Section 2 holds the number of cells t as fixed, assumes the cells probabilities in $\underset{\sim}{\pi}$ are fixed and positive, and lets the sample size N tend to infinity. Thus all of the expected cell values become large as N grows in size. At the end of Section 3 we suggested the usefulness of working with tables with an average cell size of 4 or 5. This suggests a different form of asymptotics which allows t to grow at the same rate as N while the ratio N/t is kept fixed. In this new asymptotics we replace the fixed probability $\underset{\sim}{\pi}$ by a sequence of probability vectors, lying in progressively larger probability spaces.

The emergence of interest in asymptotics of large sparse multinomials is described for the estimation of cell probabilities by Bishop, Fienberg, and Holland (1975, Chapter 12), and in context of central limit theorems

and testing by Morris (1975). Two very recent sets of results by Haberman (1977) and Koehler (1977) are of special interest in the context of the use of goodness-of-fit statistics.

6.1. Asymptotic Normality of Goodness-of-fit Statistics

The notation for the large sparse multinomial asymptotics is quite similar to that of Section 2, but we need to keep in mind the fact that the dimension and the structure of the probability and sample spaces are changing as the number of cells t grows. To do this formally we shall index the dimension of the spaces t, the sample size N, and the vectors $\underset{\sim}{X}$ and $\underset{\sim}{\Pi}$ with the subscript k, i.e. t_k, N_k, $\underset{\sim}{X}_k$ and $\underset{\sim}{\Pi}_k$, where we shall let $k \to \infty$. Strictly speaking each of the components of $\underset{\sim}{X}_k$ and $\underset{\sim}{\Pi}_k$ should also be indexed by k but such notation is for our purposes here unnecessarily elaborate.

Next, we need to define, corresponding to $\underset{\sim}{X}_k$, a vector of independent Poisson random variables $\underset{\sim}{Y}_k = (Y_1, Y_2, \ldots, Y_{t_k})$ such that

$$E(Y_i) = N_k \Pi_i = m_i \ . \tag{6.1}$$

These Poisson variates play a crucial role in describing the asymptotic structure through moments of their Kullback-Liebler information kernel

$$\begin{aligned} I(y,m) &= y \log\left(\frac{y}{m}\right) - y + m \qquad \text{for } y>0,\ m>0 \\ &= m \qquad\qquad\qquad\qquad\quad \text{for } y=0,\ m>0 \ . \end{aligned} \tag{6.2}$$

Koehler's results to date extend to loglinear models with closed-form MLE's for the elementary expected frequencies $\underset{\sim}{m}$ in both null and non-null situations, but the special case of independence in a two-dimensional table will suffice to illustrate the nature of the asymptotic structure. We

consider a sequence of tables with I_k rows and J_k columns where $\underset{\sim}{X}_k$, $\underset{\sim}{\Pi}_k$, and $\underset{\sim}{Y}_k$ are now doubly subscripted. The likelihood ratio statistic for the model of independence is

$$G_k^2 = 2 \sum_i \sum_j x_{ij} \log \left(\frac{n_k X_{ij}}{X_{i+} X_{+j}} \right) \quad , \tag{6.3}$$

and we define

$$\gamma_k = \frac{1}{N_k} \sum_i \sum_j \mathrm{Cov}[I(Y_{ij}, m_{ij}), Y_{ij}] \quad , \tag{6.4}$$

$$\sigma_k^2 = 2 \sum_i \sum_j \mathrm{Var}[I(Y_{ij}, m_{ij}) - \gamma_k Y_{ij}] \quad , \tag{6.5}$$

and

$$\begin{aligned} u_k &= 2 \sum_i \sum_j E[I(Y_{ij}, m_{ij})] - 2 \sum_i E[I(Y_{i+}, m_{i+})] \\ &\qquad - 2 \sum_j E[I(Y_{+j}, m_{+j})] \\ &= 2 \sum_i \sum_j E[Y_{ij} \log \left(\frac{N_k Y_{ij}}{Y_{i+} Y_{+j}} \right)] \end{aligned} \tag{6.6}$$

Then the distribution of G_k^2 under the null hypothesis in the large sparse asymptotics is normal in the following sense.

Theorem 8. (Koehler, 1977). Let $I_k \to \infty$ and $J_k \to \infty$ as $k \to \infty$. Suppose

$$\max_i \Pi_{i+} = 0(1) \quad \text{as} \quad I_k \to \infty \quad , \tag{6.7}$$

$$\max_j \Pi_{+j} = 0(1) \quad \text{as} \quad J_k \to \infty \quad , \tag{6.8}$$

and that there exists a fixed $\varepsilon > 0$ such that

$$N_k \Pi_{ij} = m_{ij} > \epsilon \quad \text{for all } i,j, \text{ \& } k \quad . \tag{6.9}$$

Then, if independence holds, as $k \to \infty$

$$\frac{G_k^2 - u_k}{\sigma_k} \longrightarrow z \quad , \tag{6.10}$$

where z is a normal random variable with mean 0 and variance 1.

Theorem 8 assumes that both dimensions of the table are growing in size in such a way that no marginal probability remains too large (i.e. conditions (6.7) and (6.8)) but no expected value is getting too small (i.e. condition (6.9)). The null mean of G_k^2 in the large sparse asymptotics, u_k, is not equal to the "usual" degrees of freedom, $(I_k - 1)(J_k - 1)$. Moreover u_k depends on the unknown parameters, $m_{ij} = m_{i+}m_{+j}/N_k$, and the asymptotic "suitability" of the usual MLE, $\hat{m}_{ij} = X_{i+}X_{+j}/N_k$, is somewhat questionable based on numerical results of Koehler.

The results of Williams (1976) discussed in Section 3.2 involved the approximation of the mean u_k in the usual asymptotics, using the Taylor series approximation

$$Y \log Y = m \log m + \frac{1}{2} + \frac{1}{12m} + O(m^{-2}) \tag{6.11}$$

Unfortunately, in the asymptotics of large sparse multinomials $t_k \to \infty$ and $N_k \to \infty$ but at least some of the individual cell expectations remain "moderate" in size. Exactly how this affects the suitability of the adjustment to G^2 described in Section 3.2 remains unclear, and further work needs to be done on this problem.

6.2. Likelihood Ratio Statistics for Comparing Models

Haberman (1977) has looked at a closely related problem for comparing

two loglinear models in large sparse situations. His approach is first to establish asymptotic normality of linear functionals of MLE's of log-expected values. The conditions under which his results hold are quite complex but they essentially require that elements of $\underset{\sim}{m}$ remain relative to a special norm of $\underset{\sim}{m}$, and they appear to include the conditions considered by Koehler.

Following from these results for linear functionals Haberman goes on to look at the Pearson and likelihood ratio statistics for comparing two loglinear models where one is a special case of the other. The likelihood ratio statistic is simply a difference of two statistics of the form G^2 given by (2.7), one for each of the models. In the large sparse multinomial asymptotics this amounts to considering two sequences of models (since $\underset{\sim}{m}_k$ is growing in dimension) where the difference in the estimation spaces converges to a fixed number of degrees of freedom $0 < \nu < \infty$ as $k \to \infty$. Haberman then shows that under suitable conditions that the distribution of both test statistics for comparing the fit of the two nested models converges to the usual χ^2_ν distribution.

The implications of this result for statistical practice are quite important. While the behavior of G^2 in large sparse multinomial structures requires serious attention as we saw in Section 6.1, if our primary interest in a large sparse table is focussed on the importance of a restricted subset of loglinear model parameters, Haberman's result suggests that the test statistics for comparing two models that differ by these parameters can be used with the usual χ^2 reference distributions.

Acknowledgements

This research was supported in part by Grant NIE-G76-0094 from the National Institute of Education, U.S. Department of Health, Education and Welfare, and by a grant from the Science Research Council to the Department of Statistics, University of Newcastle, where I was a visitor during the preparation of this paper. Stephen Brier, David Hinkley, Kenneth Koehler, Kinley Larntz, and Robin Plackett provided helpful comments and copies of unpublished manuscripts. An earlier version of this paper was read before the Royal Statistical Society Multivariate Study Group on March 7, 1969 at Birkbeck College, London.

References

Altham, P.M.E. (1976). "Discrete variable analysis for individuals grouped into families." Biometrika 63, 263-269.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Brier, S.S. (1978). Categorical Data Models for Complex Data Structures. Unpublished Ph.D. dissertation, School of Statistics, University of Minnesota.

Chase, G.R. (1972). "On the chi-square test when the parameters are estimated independently of the sample." J. Amer. Statist. Assoc. 67, 609-611.

Cochran, W.G. (1952). "The χ^2 test of goodness of fit." Ann. Math. Statist. 23, 315-346.

Cohen, J.C. (1976). "The distribution of the chi-squared statistic under clustered sampling from contingency tables. J. Amer. Statist. Assoc. 71, 665-670.

Finney, D.J. (1971). Probit Analysis (3rd edition). Cambridge University Press.

Fisher, R.A. (1958). Statistical Methods for Research Workers (13th Edition). London: Oliver and Boyd.

Haberman, S.J. (1977). "Log-linear models and frequency tables with small expected cell counts." Ann. Statist. 5, 1148-1169.

Koehler, K.J. (1977). Goodness of Fit Statistics for Large Sparse Multinomials. Unpublished Ph.D. dissertation, School of Statistics, University of Minnesota.

Larntz, K. (1971). Some Models for Individual-Group Behavior and Group Comparisons. Unpublished Ph.D. dissertation, Department of Statistics, University of Chicago.

Larntz, K. (1978). "Small sample comparisons of exact levels for chi-square goodness of fit statistics." J. Amer. Statist. Assoc. (in press).

Morris, C. (1975). "Central limit theorems for multinomial sums." Ann. Statist. 3, 165-188.

Mosimann, J.E. (1962). "On the compound multinomial distribution, the multivariate β-distribution, and correlation among proportions." Biometrika 49, 65-82.

Murthy, V.K. and Gafarian, A.V. (1970). "Limiting distributions of some variations of the chi-square statistics." Ann. Math. Statist. 41 188-194.

Odoroff, C.L. (1970). "A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2 x 2 x 2 and 3 x 2 x 2 contingency tables: tests for interaction," J. Amer. Statist. Assoc. 65, 1617-1631.

Plackett, R.L. (1974). The Analysis of Categorical Data. London: Griffin.

Plackett, R.L. and Paul, S.R. (1978a). "Dirichilet models for square contingency tables." Communications in Statistics (to appear).

Plackett, R.L. and Paul, S.R. (1978b). "Inference sensitivity for Poisson mixtures." Unpublished manuscript.

Stone, M. (1974). "Cross validation and multinomial prediction." Biometrika 61, 509-515.

Watson, G.S. (1959). "Some recent results on chi-square goodness-of-fit tests." Biometrics 15, 440-468.

Williams, D.A. (1976). "Improved likelihood ratio tests for complete contingency tables." Biometrika 63, 33-37.

Yarnold, J.K. (1970). "The minimum expectation of χ^2 goodness-of-fit tests and the accuracy of approximations for the null distribution." J. Ameri. Statist. Assoc. 65, 864-886.

1. REJECTION RATES FOR 3×3×3 NO THREE-FACTOR INTERACTION MODEL (SOURCE: LARNTZ, 1978).

Row Margins Proportional to:	Column Margins Proportional to:	Layer Margins Proportional to:		Sample Size 20	40	60	80	100
2:3:5	2:3:5	2:3:5	X^2	.0175	.0550	.0585	.0645	.0690
			G^2	.0190	.0885	.1125	.1160	.1265
			T^2	.0010	.0175	.0475	.0600	.0775
2:3:5	2:3:5	6:6:7	X^2	.0435	.0820	.0650	.0740	.0485
			G^2	.0335	.1220	.1375	.1340	.0925
			T^2	.0025	.0370	.0635	.0835	.0630
2:3:5	6:6:7	6:6:7	X^2	.0440	.0710	.0675	.0700	.0565
			G^2	.0575	.1410	.1475	.1275	.1035
			T^2	.0025	.0400	.0665	.0875	.0725
6:6:7	6:6:7	6:6:7	X^2	.0825	.0870	.0855	.0625	.0680
			G^2	.0950	.1740	.1660	.1140	.1120
			T^2	.0045	.0820	.1085	.0805	.0840

NOTE: Values are based on 2,000 trials with the same trials used for X^2, G^2, and T^2. Approximate standard error (based on true level of .05) for each value is .0049.